Introduction to Data Science

BRIAN D'ALESSANDRO

ADJUNCT PROFESSOR, NYU

FALL 2018

<u>Fine Print</u>: *these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute, except for as needed as a pedagogical tool in the subject of Data Science.*

DATA MINING PROCESS OVERVIEW

NYU – Intro to Data Science

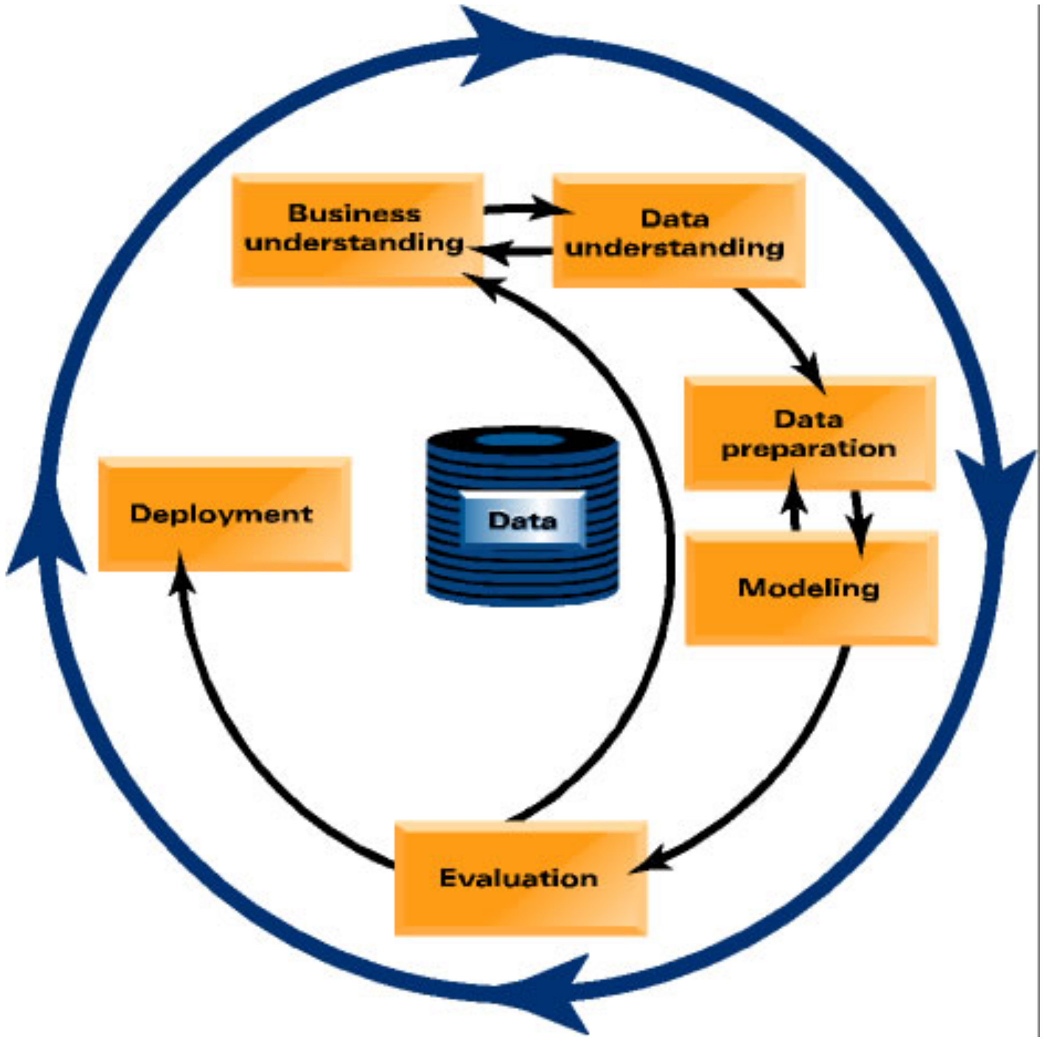
FUNDAMENTAL CONCEPTS

- Extracting useful knowledge from data to solve business problems can be treated systematically by following a process with reasonably well-defined stages.

- Formulating data mining solutions and evaluating the results involves thinking carefully about the context in which they will be used.

DATA MINING PROCESS

Cross Industry Standard Process for Data Mining



WHY EMPHASIZE PROCESS

1. **Rigor** – each stage in the process is generally supported by well researched theoretical principles or well documented heuristics

WHY EMPHASIZE PROCESS

1. **Rigor** – each stage in the process is generally supported by well researched theoretical principles or well documented heuristics

2. **Reliability** – data mining tasks stand up better to peer and managerial review when the tasks adhere to common process fundamental principles and standards

WHY EMPHASIZE PROCESS

1. **Rigor** – each stage in the process is generally supported by well researched theoretical principles or well documented heuristics

2. **Reliability** – data mining tasks stand up better to peer and managerial review when the tasks adhere to common process fundamental principles and standards

3. **Reproducibility** – with a well defined process we can better replicate results and also automate certain learning tasks

COMMON BUSINESS QUESTIONS

The following lists common questions we face in industrial settings that can be addressed using data and the tools of data science.

- Will customer *X* churn next month/default on her loan?
- How much would prospect *X* spend if they were a customer?
- Who might be good "friends" on our social networking site?
- Did *X* cause *Y* to happen?
- What should you recommend to user *I*.
- Do users fall into unique groups?
- Is this transaction fraudulent?

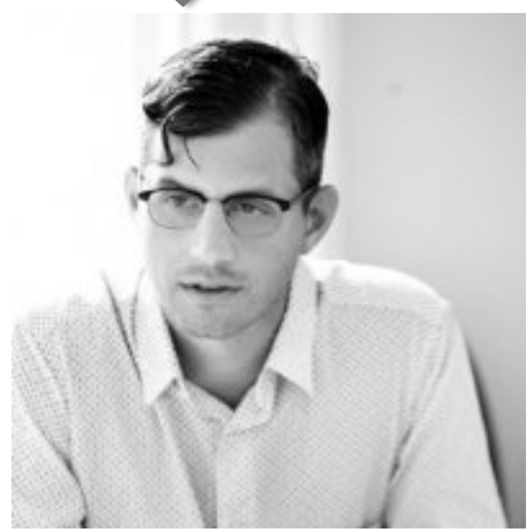
PROBLEM FORMULATION -TRANSLATION

Data Scientists speak a different language, and you need to be able to translate. This means formulating business objectives in the language of data science.

Let me test whether or not adding incremental data assets improves the lift of our models. I can then measure the net economic benefit and normalize by cost.

We should invest in more data, but only if it drives positive ROI!

CEO



Data Scientist

PROJECT/BUSINESS UNDERSTANDING

Put the problem into context…ask questions…be creative!

Be prepared to ask…

- What is the goal of the solution?
- Why do we need to do this?
- What data is available?
- What constraints exist?
- What is an acceptable solution?
- How do we measure?
- What is success?

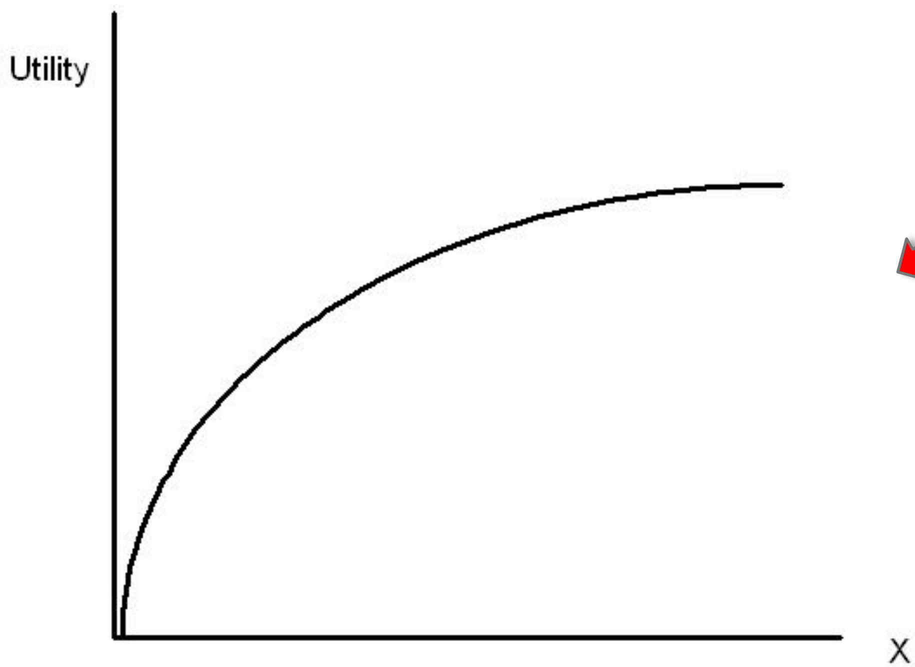
→

- Exec
- Sales
- Marketing
- Technology
- Operations

TIPS FOR PROBLEM FORMULATION

Simplify the problem and iterate as much as possible

Keep the problem simpler at first, add more to it later.



Model Complexity and Effort Building/Implementing

A good but simple model is always better than no model!

Bias yourself towards deployment when competing against time.

DATA

Clearly the most important topic yet…

Rules of thumb

1. Know where your data comes from.

2. Know how to get the data.

3. Know what your data looks like.

4. Know the limits of your data.

Don’t worry, we will cover this topic extensively!

IDEAL SCIENTIFIC METHOD

The original intention of data was to falsify/confirm a hypothesis about the world.

1. Make a hypothesis

2. Collect data

3. Falsify/confirm hypothesis

ANALYSIS IN A 'BIG DATA' WORLD

The ubiquity of data comes with a catch – spurious signal is everywhere. Ignoring the core tenets of the scientific method can lead to bad results and even harm. Don't make this mistake…

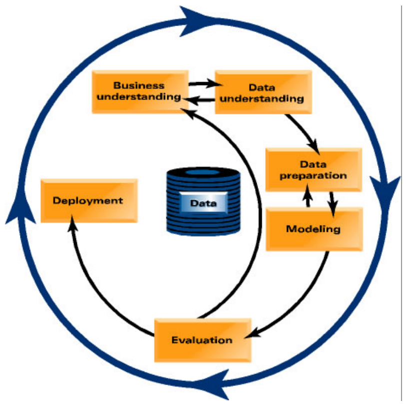
1. Look at data lying around

2. Make a hypothesis

3. Falsify/confirm hypothesis on same data

4. Make decisions that lead to poor generalization

FOOTNOTE: RESOLVING THE CIRCULAR NATURE OF ANALYSIS



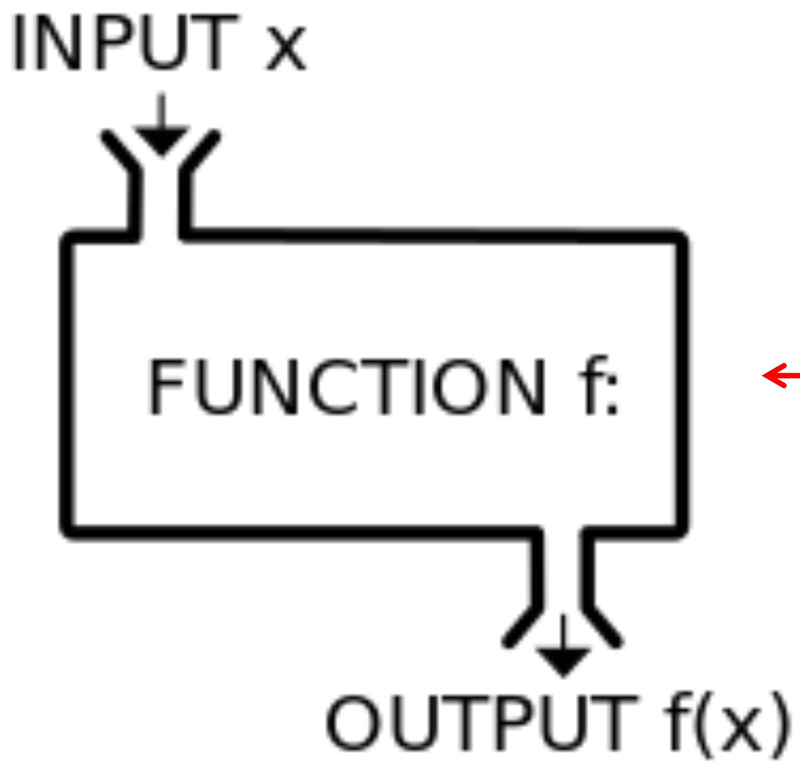
In reality, you will always likely have looked at some data before analyzing some hypothesis.

The solution is to make every effort to bring in fresh data or use a holdout set upon each iteration of testing/modeling.

We will cover this concept in depth in a supervised learning context.

MODELING

The engine of data science.



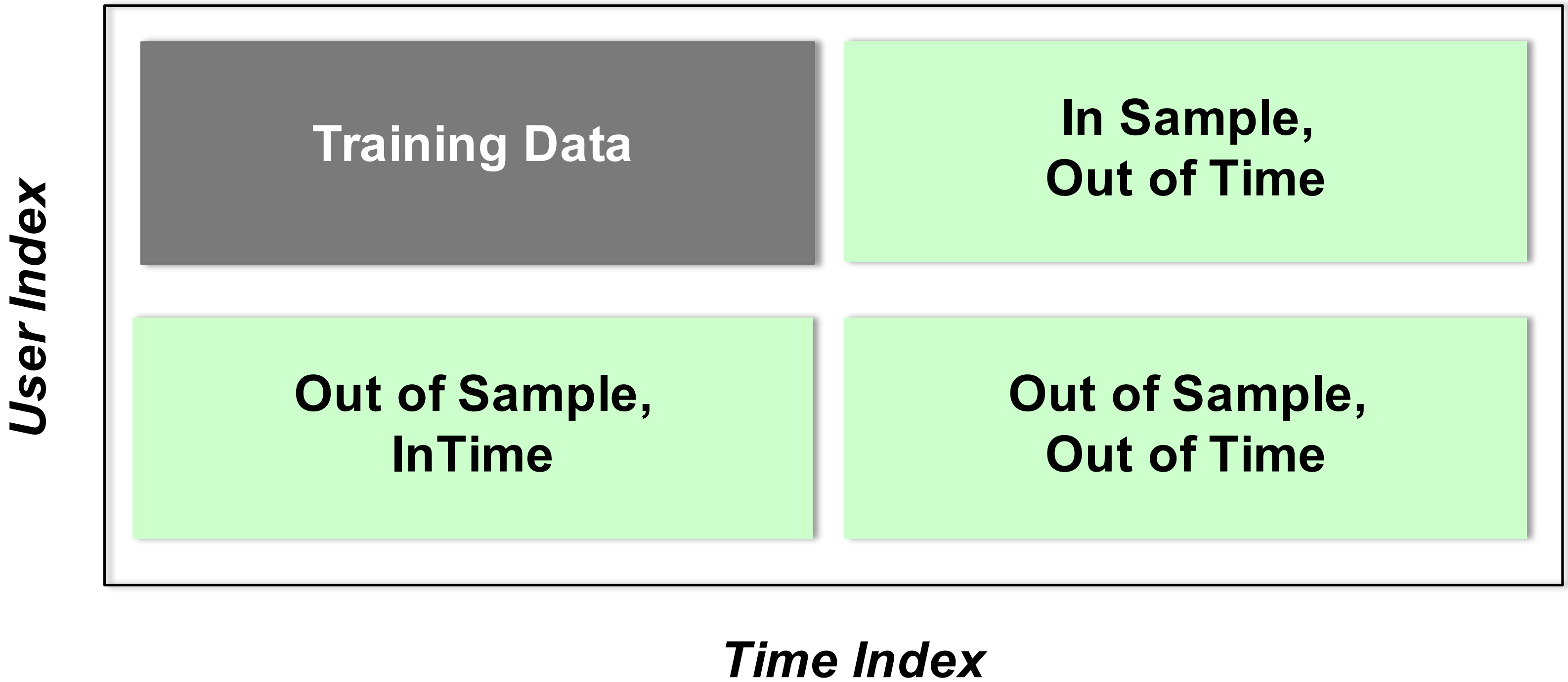
Modeling is how you get from data to insights and decision making.

We will cover how this is done extensively in this course.

EVALUATION

The safety net of data science.

Evaluation should be built in automatically to the modeling process.



Throughout this class we will learn various evaluation methodologies along with some of the theory as to why proper evaluation is critically important.

DEPLOYMENT

Your model and analysis are nothing without action.

When your model is shipped to a production system:

- Don't walk away – your model isn't what you think it is, its what the developer thinks it is.
- You are the steward and caretaker. Be proactive about QA and regular performance monitoring.

When your analysis is delivered to people

- Communication is everything
- Use data to tell a story
- Connect your analysis to the audiences' goals
- Collect feedback

FULL CIRCLE

Once deployed, its not over. Start thinking about the next iteration!